Table of Contents

1. Introduction

The project aims to detect Alzheimer's disease based on demographic, medical, and clinical features. Alzheimer's is a progressive neurological disorder and the most common cause of dementia. Early detection can help manage and treat the disease better. Therefore our goal is to build a model that can accurately predict Alzheimer's for a data-driven classification. Our motivation comes from building an ML model that classifies Alzheimer's before it turns to the terminal stage so that it can be cured or managed.

---

2. Dataset Description

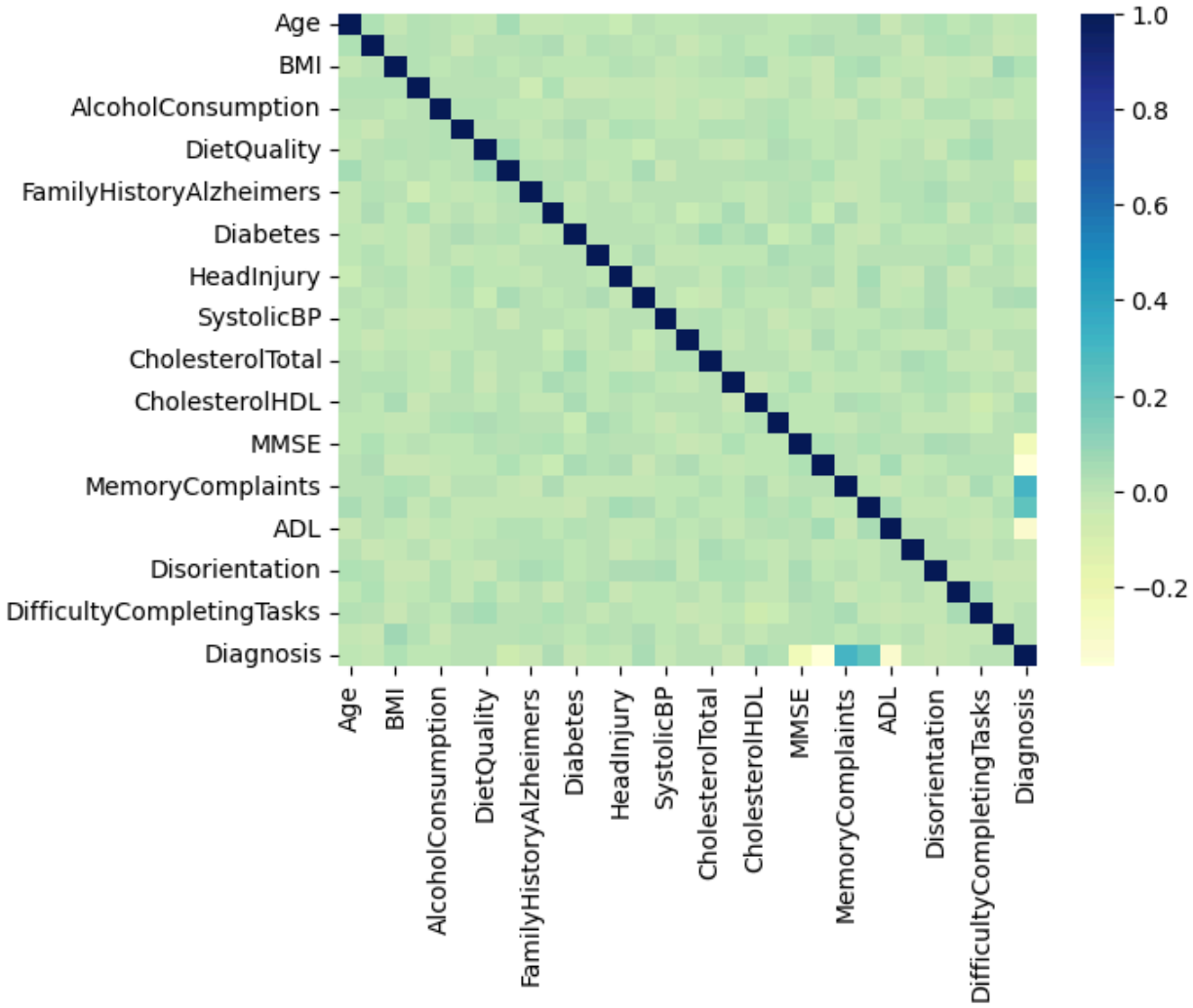
Source:

- **Link:** https://www.kaggle.com/dsv/8668279
- **Reference:** Rabie El Kharoua. (2024). *Alzheimer's Disease Dataset*. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/8668279

Dataset Description:

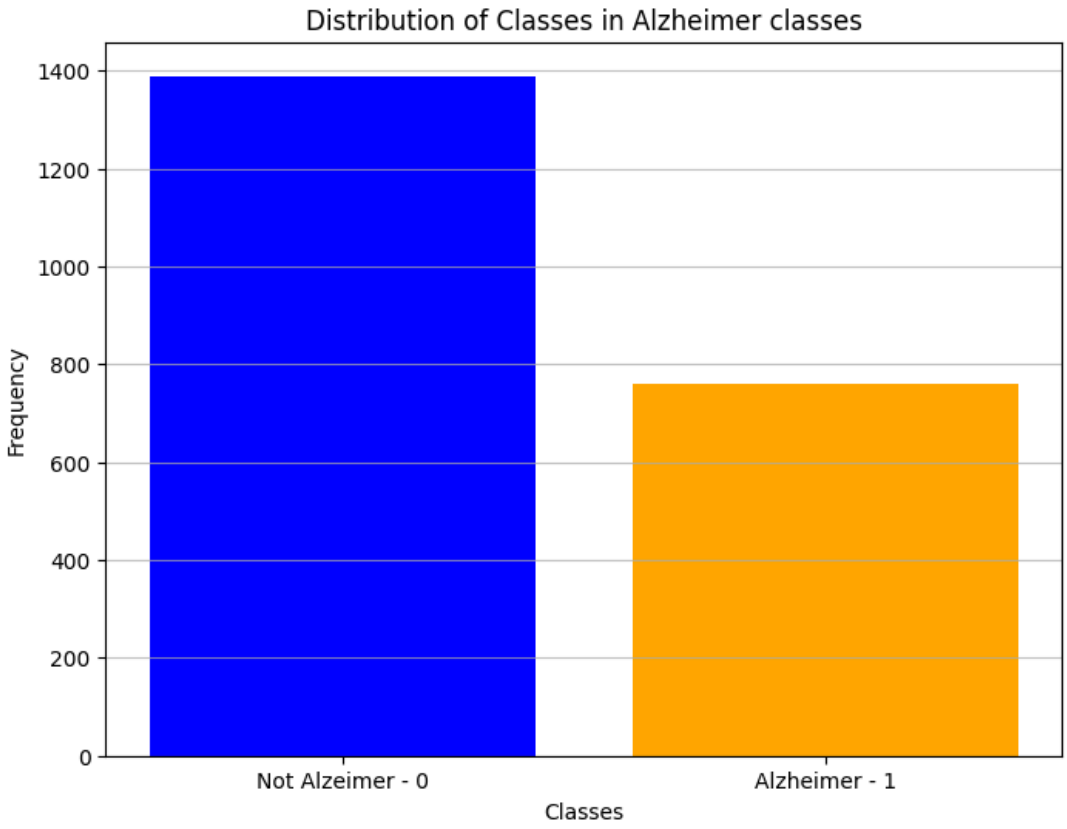
- **Number of Features:** 30 (after removing irrelevant features PatientID, Ethnicity, EducationLevel, and DoctorInCharge)
- **Problem Type:** Classification Problem
  - The goal is to predict a binary output (1: Alzheimer's, 0: Not Alzheimer's).
- **Number of Data Points:** 2149 patients
- **Feature Types:**
  - **Quantitative:** Age, BMI, AlcoholConsumption, PhysicalActivity DietQuality, SleepQuality, SystolicBP, DiastolicBP, CholesterolTotal, cholesterol LDL, CholesterolHDL, CholesterolTriglycerides, MMSE, FunctionalAssessment, ADL
  - **Categorical:** Gender, Smoking status, FamilyHistoryAlzheimers, cardiovascular disease, Diabetes, Depression, HeadInjury, Hypertension, MemoryComplaints, BehavioralProblems, Confusion, Disorientation PersonalityChanges, DifficultyCompletingTasks, Forgetfulness.

Correlation: Using Seaborn Library we can see there aren't any major correlations between the features.



Imbalanced Dataset: The dataset is imbalanced.

- Instances of class 1 (Alzheimer's): 760
- Instances of class 0 (Not Alzheimer's): 1389



3. Dataset Pre-processing

Fault	**Solution Implemented**
Irrelevant features present such as PatientID, Ethnicity, EducationLevel, and DoctorInCharge. Which do not affect the outcome.	Removed irrelevant columns (PatientID, Ethnicity, EducationLevel, and DoctorInCharge).
There were categorical features.	All of them were Binary categorical variables (e.g., Smoking, Gender). Thus were already mapped to 0 and 1. Ethnicity was the only categorical value that needed imputing but it was dropped.
No NULL values were detected.	No need for data imputing.

---

4. Feature Scaling

- StandardScaler and MinMaxScaler were used to compare the accuracy using KNN.
- Without any scaling, the accuracy was 0.58
- MinMaxScaler Accuracy was 0.70
- StandardScaler Accuracy was 0.75

  Thus we used StandardScaler for feature scaling

---

5. Dataset Splitting

- Data was split randomly:
  - Training Set: 70%
  - Testing Set: 30%
  - Random state 42

---

6. Model Training & Testing

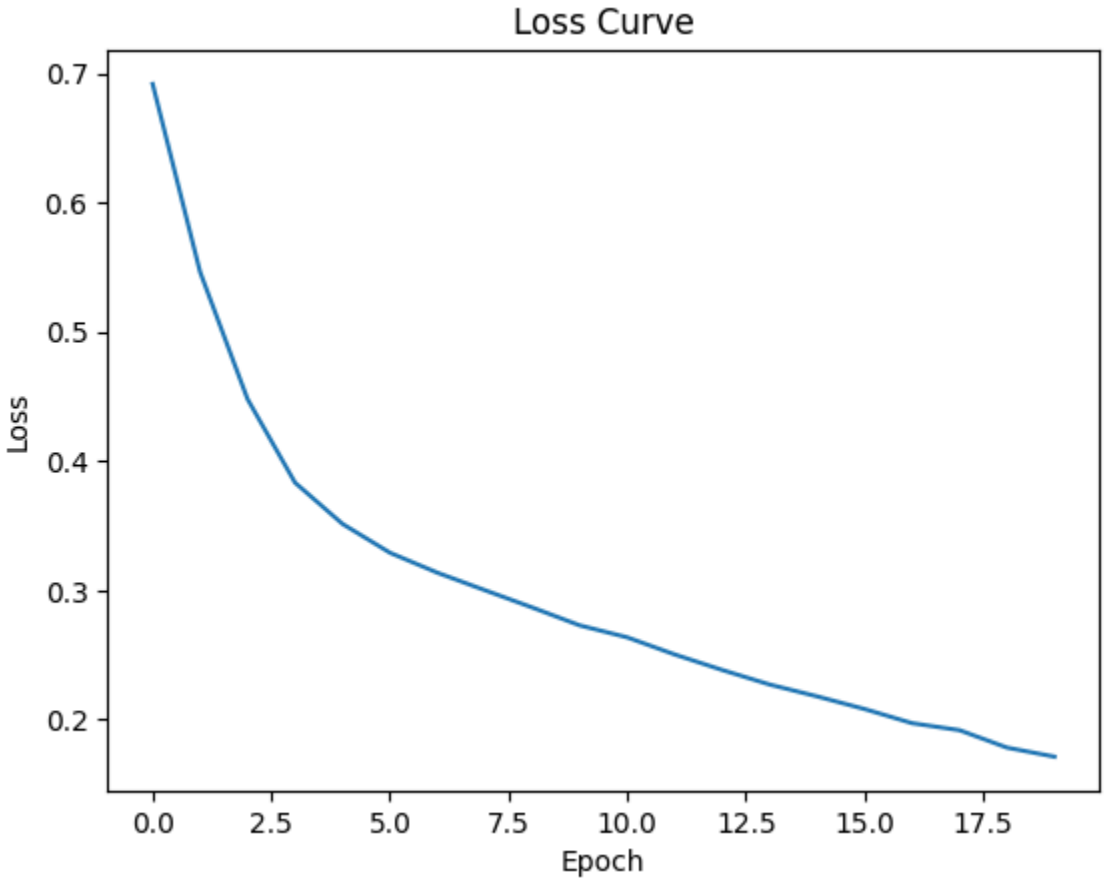
Algorithms Applied:

As our dataset is supervised and the output class is discrete categorical (0 or 1)

Thus we used the following algorithms to evaluate the outcome.

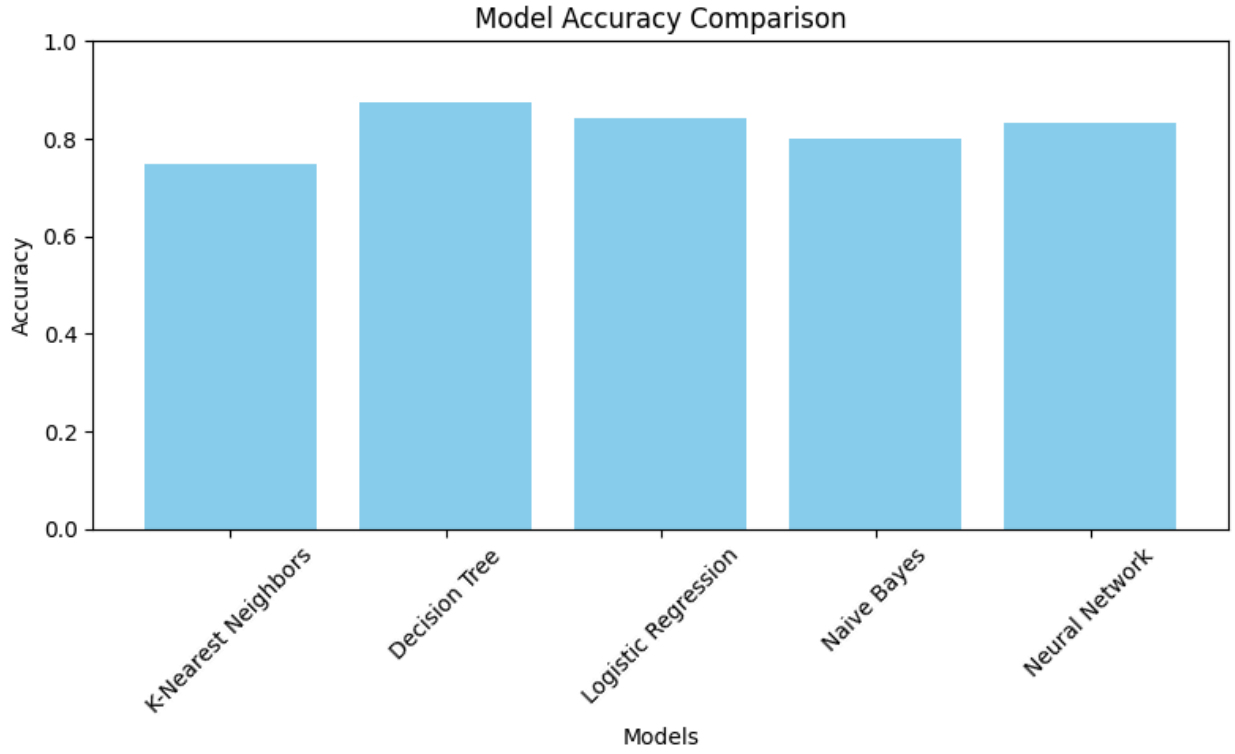
1. **K-Nearest Neighbors (KNN):**
   - Parameters: Default settings
2. **Decision Tree:**
   - Parameters: Random state set to 42
3. **Logistic Regression:**
   - Parameters: Random state set to 42
4. **Gaussian Naive Bias:**
   - Parameters: Default settings

5. **Neural Network:**
    - Architecture: 3-layer feed-forward neural network
    - Activation function for the hidden layers: Relu
    - Activation function for the output layer: Sigmoid
    - Loss function: Binary Cross-Entropy
    - Metrics: Loss curve, confusion matrix, and classification report were included.



---

7. Model Selection/Comparison Analysis



Model	Accuracy
KNN	74.7%
Decision Tree	87.44%
Logistic Regression	84.18%
Naive Bias	79.8%
Neural Network	83.1%

TP, FN, TN, FP, Precision, Recall & F1 Score Table

Model	TP	FN	TN	FP	Precision	Recall	F1 Score
KNN	369	115	113	48	0.74	0.75	0.73
Decision Tree	367	31	197	50	0.88	0.87	0.88
Logestic Regression	369	54	174	48	0.84	0.84	.84
Naive Bias	350	63	165	67	0.8	0.8	0.8
Neural Network	362	54	174	55	0.83	0.83	0.83

8. Conclusion

The project classified Alzheimer's disease using multiple machine-learning models. Among the models, the Decision Tree provided the highest accuracy at 87.44%, followed by the Logistic Regression at 84.18%. Neural Network also performed well. Using more complex architectures I can achieve better results. This project is a crucial demonstration of data-driven classification to detect Alzheimer's at an early stage based on lifestyle and clinical data which is very significant in saving many people's lives.